

**Scholar All articles - Recent articles** Results 1 - 10 of about 2,550 for **duplicate document webcrawler OR crawler** (0.17 seconds)**Mercator: A scalable, extensible web crawler-** [unwo.ca](#) [PDF]

A Heydon, M Najork - World Wide Web, 1999 - Springer

... a **document** more than once, a Web **crawler** may wish ... down- loaded **documents** that are **duplicates** of pages ... saved the complete contents of every downloaded **document** ...

Cited by 269 - Related articles - Web Search - BL Direct - All 49 versions

**Identifying and filtering near-duplicate documents-** [princeton.edu](#) [PDF]

AZ Broder - Lecture notes in computer science, 2000 - Springer

... and "robot traps"), and erroneously (**crawler** or server ... main reasons: first, indexing of **duplicates** wastes ex ... seldom interested in seeing **documents** that are ...

Cited by 81 - Related articles - Web Search - BL Direct - All 6 versions

**OverCite: A cooperative digital research library-** [psu.edu](#) [PDF]

J Stribling, IG Council, J Li, MF Kaashoek, DR ... - Lecture notes in computer science, 2005 - Springer

... download the file. After the download, the **crawler** process checks whether this is a **duplicate document**. This requires (1) looking ...

Cited by 47 - Related articles - Web Search - Library Search - BL Direct - All 38 versions

**[PDF] -On the evolution of clusters of near-duplicate web pages**

D Fetterly, M Manasse, M Najork - Proceedings of the 1st Latin American Web Congress, 2003 - cwr.cl

... that have been found to be near-**duplicates** of one ... the data using the Mercator web **crawler** [12], customized ... by whitespace, and then segmented the **document** into 5 ...

Cited by 62 - Related articles - View as HTML - Web Search - All 21 versions

**Design and implementation of a distributed crawler and filtering processor**

D Zeinalipour-Yazdi, M Dikaiakos - Lecture notes in computer science, 2002 - Springer

... executed repeatedly until all links of the **document** at hand ... added to the URL-Queue, dropping all **duplicate** URL's ... that have been visited by the **crawler** already ...

Cited by 21 - Related articles - Web Search - BL Direct - All 7 versions

**[PDF] -Large linguistically-processed Web corpora for multiple languages**

M Baroni, A Kilgarriff - Proceedings of European ACL, 2006 - acl.ldc.upenn.edu

... The crawls are performed using the Her- itrix **crawler**, 4 with a ... at least one duplicate, we discard not only the **duplicate(s)** but also the **document** itself ...

Cited by 30 - Related articles - View as HTML - Web Search - All 14 versions

**Information fusion with ProFusion**

S Gauch, G Wang - ACM-SIGIR97--Workshop on Networked Information Retrieval, 1997 - ad.informatik.uni-freiburg.de

... Excite, InfoSeek, Lycos, Open Text, **WebCrawler**); ProFusion; and two ... the number of irrelevant **documents**, the number ... links, the number of **duplicates**, the number ...

Cited by 39 - Related articles - Cached - Web Search - All 16 versions

**ProFusion\*: Intelligent fusion from multiple, distributed search engines-** [iukm.org](#)

S Gauch, G Wang, M Gomez - Journal of Universal Computer Science, 1996 - jucs.org

... Excite, InfoSeek, Lycos, Open Text, **WebCrawler**); ProFusion; and two ... the number of irrelevant **documents**, the number ... links, the number of **duplicates**, the number ...

Cited by 138 - Related articles - Web Search - All 20 versions

## Web information retrieval-an algorithmic perspective

M Henzinger - Lecture notes in computer science, 2000 - Springer

... The indexer processes the pages collected by the **crawler**. First it decides which of them to index. For example, it might discard **duplicate documents**. ...

Cited by 14 - Related articles - Web Search - DL Direct - All 7 versions

## Finding near-duplicate web pages: a large-scale evaluation of algorithms- +chinaunix.net (PDF)

M Henzinger - Proceedings of the 29th annual international ACM SIGIR ..., 2006 - portal.acm.org

... it uses the same amount of space per **document** and returns ... collected during a crawl of Google's **crawler**. ... and 2.2% of the pages after **duplicate** removal have ...

Cited by 50 - Related articles - Web Search - All 13 versions

Key authors: M Najork - A Heydon - E Selberg - S Gauch - O Etzioni

Google ►

Result Page: 1 2 3 4 5 6 7 8 9 10

[Next](#)

[Go to Google Home](#) - [About Google](#) - [About Google Scholar](#)

©2009 Google